OpenBias: Open-set Bias Detection in Text-to-Image Generative Models

Moreno D'Incà[1], Elia Peruzzo[1], Massimiliano Mancini[1], Dejia Xu[2], Vidit Goel[3,4], Xingqian Xu[3,4], Zhangyang Wang[2,4], Humphrey Shi[3,4†], Nicu Sebe[1†]

[1]University of Trento, [2]UT Austin, [3]SHI Labs @ Georgia Tech & UIUC, [4]Picsart AI Research (PAIR)

https://github.com/Picsart-AI-Research/OpenBias

Abstract

Text-to-image generative models are becoming increasingly popular and accessible to the general public. As these models see large-scale deployments, it is necessary to deeply investigate their safety and fairness to not disseminate and perpetuate any kind of biases. However, existing works focus on detecting closed sets of biases defined a priori, limiting the studies to well-known concepts. In this paper, we tackle the challenge of open-set bias detection in text-to-image generative models presenting OpenBias, a new pipeline that identifies and quantifies the severity of biases agnostically, without access to any precompiled set. OpenBias has three stages. In the first phase, we leverage a Large Language Model (LLM) to propose biases given a set of captions. Secondly, the target generative model produces images using the same set of captions. Lastly, a Vision Question Answering model recognizes the presence and extent of the previously proposed biases. We study the behavior of Stable Diffusion 1.5, 2, and XL emphasizing new biases, never investigated before. Via quantitative experiments, we demonstrate that OpenBias agrees with current closed-set bias detection methods and human judgement.

1. Introduction

Text-to-Image (T2I) generation has become increasingly popular, thanks to its intuitive conditioning and the high quality and fidelity of the generated content [39, 41, 43, 44, 46]. Several works extended the base T2I model, unlocking additional use cases, including personalization [16, 45], image editing [7, 14, 18, 21], and various forms of conditioning [2, 24, 63]. This rapid progress urges to investigate other key aspects beyond image quality improvements, such as their fairness and potential bias perpetration [11, 15, 62]. It is widely acknowledged that deep learning models learn the underlying biases present in their training sets [3, 20, 64], and generative models are no exception [11, 15, 36, 62].

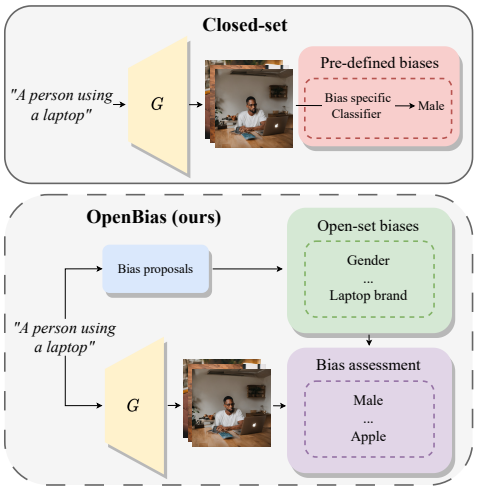

Figure 1. OpenBias discovers biases in T2I models within an open-set scenario. In contrast to previous works [15, 29, 62], our pipeline does not require a predefined list of biases but proposes a set of novel domain-specific biases.

Ethical topics such as fairness and biases have seen many definitions and frameworks [55]; defining them comprehensively poses a challenge, as interpretations vary and are subjective to the individual user. Following previous works [15, 60], a model is considered unbiased regarding a specific concept if, given a context t that is agnostic to class distinctions, the possible classes $c \in \mathcal{C}$ exhibit a uniform distribution. In practice, for a T2I model, this reflects to the tendency of the generator to produce content of a certain class c (*e.g. "man"*), given a textual prompt t that does not specify the intended class (*e.g. "A picture of a doctor"*).

Several works studied bias mitigation in pre-trained models, by introducing training-related methods [25, 37, 47, 57] or using data augmentation techniques [1, 12]. Nevertheless, a notable limitation of these approaches is their dependence on a predefined set of biases, such as gender, age, and race [11, 15], as well as specific face attributes [62]. While these represent perhaps the most sensitive biases, we argue that there could be biases that remain undiscovered and unstudied. Considering the example in

† Corresponding authors.

Fig.1, the prompt *"A person using a laptop"* does not specify the person's appearance and neither the specific laptop nor the scenario. While closed-set pipelines can detect well-known biases (*e.g*. gender, race), the T2I model may exhibit biases also for other elements (*e.g*. laptop brand, office). Thus, an open research question is: *Can we identify arbitrary biases present in T2I models given only prompts and no pre-specified classes?* This is challenging as collecting annotated data for all potential biases is prohibitive.

Toward this goal, we propose *OpenBias*, the first pipeline that operates in an *open-set scenario*, enabling to identify, recognize, and quantify biases in a specific T2I model without constraints (or data collection) for a specific predefined set. Specifically, we exploit the multi-modal nature of T2I models and create a knowledge base of possible biases given a collection of target textual captions, by querying a Large Language Model (LLM). In this way, we discover specific biases for the given captions. Next, we need to recognize whether these biases are actually present in the images. For this step, we leverage available Visual Question Answering (VQA) models, directly using them to assess the bias presence. By doing this, we overcome the limitation of using attributes-specific classifiers as done in previous works [15, 50, 62], which is not efficient nor feasible in an open-set scenario. Our pipeline is modular and flexible, allowing for the seamless replacement of each component with newer or domain-specific versions as they become available. Moreover, we treat the generative model as a *black box*, querying it with specific prompts to mimic end-user interactions (*i.e*. without control over training data and algorithm). We test OpenBias on variants of Stable Diffusion [41, 44] showing human-agreement, model-level comparisons, and the discovery of novel biases.

Contributions. To summarize, our key contributions are:

- To the best of our knowledge, we are the first to study the problem of open-set bias detection at large scale without relying on a predefined list of biases. Our method discovers novel biases that have never been studied before.
- We propose OpenBias, a modular pipeline, that, given a list of prompts, leverages a Large Language Model to extract a knowledge base of possible biases, and a Vision Question Answer model to recognize and quantify them.
- We test our pipeline on multiple text-to-image generative models: Stable Diffusion XL, 1.5, 2 [41, 44]. We assess our pipeline showing its agreement with closed-set classifier-based methods and with human judgement.

2. Related work

Pipeline with Foundation Models. We broadly refer to foundation models [4] as large-scale deep learning models trained on extensive data corpora, usually with a self-supervised objective [4]. This approach has been used across different modalities, such as text [8, 54], vision [9, 13, 40] and multi-modal models [35, 42, 65]. These models can be fine-tuned on downstream tasks or applied in a zero-shot manner, generalizing to unseen tasks [8, 51, 58].

Lately, several works combined different foundation models to solve complex tasks. [19, 52] use an LLM to generate Python code that invokes vision-language models to produce results. TIFA [23] assesses the faithfulness of a generated image to a given text prompt, by querying a VQA model with questions produced by an LLM from the original caption. Similarly, [10, 65] enhance image/video captioning by iteratively querying an LLM to ask questions to a VQA model. Differently, [30] identify spurious correlations in synthetic images via captioning and language interpretation, but without categorizing or quantifying bias.

We share a similar motivation, i.e., we leverage powerful foundation models to build an automatic pipeline, tailored to the novel task of open-set bias discovery. OpenBias builds a knowledge base of biases leveraging the domain-specific knowledge from *real* captions and LLMs.

Bias Mitigation in Generative Models. Bias mitigation is a long-studied topic in generative models. A substantial line of work focused on GAN-based methods. Some works improve fairness at inference time by altering the latent space semantic distribution [53] or by gradient clipping to control the gradient ensuring fairer representations for sensitive groups [29]. The advent of T2I generative models has directed research efforts towards fairness within this domain. FairDiffusion [15] guides Stable Diffusion [44] toward fairer generation in job-related contexts. It enhances classifier-free guidance [22] by adding a fair guidance term based on user-provided fair instructions. Similarly, [6] demonstrates that (negative) prompt and semantic guidance [5] mitigate inappropriateness generation in several T2I models. Given handwritten text as input, *ITI-GEN* [62] enhances the fairness of T2I generative models through prompt learning. To improve fairness, [50] guide generation using the data manifold of the training set, estimated via unsupervised learning.

While yielding notable result, these bias mitigation methods rely on predefined lists of biases. Here, we argue that there may exist other biases not considered by these methods. Therefore, our proposed pipeline is orthogonal, providing a valuable tool to enhance their utility.

3. OpenBias

This section presents OpenBias, our pipeline for proposing, assessing, and quantifying biases in T2I generative models. The overview of the proposed framework is outlined in Fig. 2. Starting from a dataset of real textual captions, we leverage a Large Language Model (LLM) to build a knowledge base of possible biases that may occur during image generation. This process enables the identification of domain-specific biases unexplored up to now. In the sec-

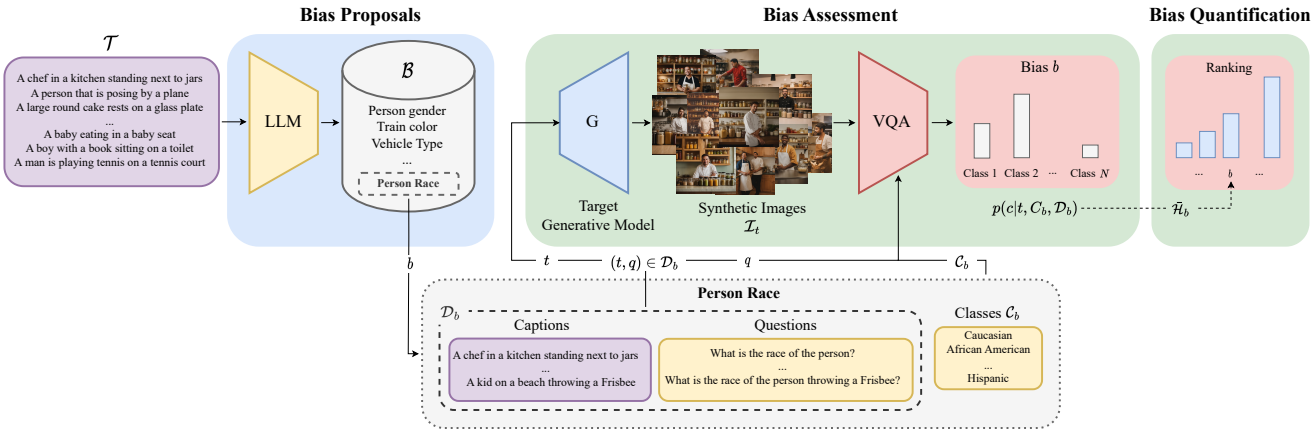

Figure 2. OpenBias pipeline. Starting with a dataset of real textual captions ($\mathcal{T}$) we leverage a Large Language Model (LLM) to build a knowledge base $\mathcal{B}$ of possible biases that may occur during the image generation process. In the second stage, synthesized images are generated using the target generative model conditioned on captions where a potential bias has been identified. Finally, the biases are assessed and quantified by querying a VQA model with caption-specific questions extracted during the bias proposal phase.

ond stage, we synthesize images using the target generative model, conditioned on captions where a potential bias has been identified. Lastly, we assess the biases with a VQA model, querying it with caption-specific questions generated during the bias proposal phase.

3.1. Bias Proposals

Given a dataset of real captions $\mathcal{T}$, we construct a knowledge base $\mathcal{B}$ of possible biases. For each caption in the dataset, we task a LLM with providing three outputs: the potential bias name, a set of classes associated with the bias, and a question to identify the bias.

Formally, given a caption $t \in \mathcal{T}$, let us denote the LLM's output as a set of triplets $\mathtt{L}_t = \{(b_i^t, \mathcal{C}_i^t, q_i^t)\}_{i=1}^{n_t}$ where the cardinality of the set n_t is caption dependent, and each triplet $(b, \mathcal{C}, q)$ has a proposed bias b, a set of associated classes $\mathcal{C}$, and the question q assigned to the specific caption t. To obtain this set, we propose to use in-context learning [8, 49], providing task description and demonstrations directly in the textual prompt.[1] We build the knowledge base $\mathcal{B}$ by aggregating the per-caption information on the whole dataset. Specifically, we can define the set of caption-specific biases B_t as the union of its potential biases, *i.e.* $B_t = \bigcup_{i=1}^{n_t} b_i$. The dataset-level set of biases is then the union of the caption-level ones, *i.e.* $\mathcal{B} = \bigcup_{t \in \mathcal{T}} B_t$. Next, we aggregate the bias-specific information across the whole dataset. We define the database of captions and questions as

$$\mathcal{D}_b = \{(t, q) \mid \forall t \in \mathcal{T}, (x, \mathcal{C}, q) \in \mathtt{L}_t, x = b\}. \quad (1)$$

$\mathcal{D}_b$ collects captions and questions specific to the bias b. Moreover, we define $\mathcal{T}_b = \{t \mid (t, q) \in \mathcal{D}_b\}$ as the set of captions, and $\mathcal{C}_b$ is the union of the set of classes associated to the bias b in $\mathcal{T}$. Nevertheless, $\mathcal{D}_b$ does not account for the potential specification of the classes of b in the caption. For instance, if we aim to generate "*An image of a large dog*", the dog's size should not be included among the biases. To address this, we implement a two-stage filtering procedure of $\mathcal{D}_b$. First, given a pair $(t, q) \in \mathcal{D}_b$ we ask the LLM to output whether the answer to the question q is explicitly present in the caption t. Secondly, we leverage ConceptNet [48] to identify synonyms for the classes $\mathcal{C}_b$ related to the specific bias b, and filter out the captions in containing either a class $\mathcal{C}_b$ or its synonyms. We empirically observe that combining these two stages produces more robust results.

By executing the aforementioned steps, we generate bias proposals in an open-set manner tailored to the given dataset. In the following sections, we elaborate on the process of bias quantification in a target generative model.

3.2. Bias Assessment and Quantification

Let G be the target T2I generative model. Our objective is to evaluate if G generates images with the identified biases. Given a bias $b \in \mathcal{B}$ and a caption $t \in \mathcal{T}_b$, we generate the set of N images $\mathcal{I}_b^t$ as

$$\mathcal{I}_b^t = \{\mathrm{G}(t, s) | \forall s \in S\} \quad (2)$$

where S is the set of sampled random noise, of cardinality $|S| = N$. Sampling multiple noise vectors allows us to obtain a distribution of the G output on the same prompt t.

To assess the bias within $\mathcal{I}_b^t$, we propose to leverage a state-of-the-art Vision Question Answering (VQA) model `VQA` mapping images and questions to answers in natural language. The VQA processes the images $\mathcal{I}_b^t$, and their associated question q in the pair $(t, q) \in \mathcal{D}_b$, choosing an answer from the possible classes $\mathcal{C}_b$. Formally, given an image

[1] We refer the reader to the *Supp. Mat.* for system prompt details.

$I \in \mathcal{I}_b^t$ we denote the predicted class as

$$\hat{c} = \texttt{VQA}(I, q, \mathcal{C}_b). \tag{3}$$

With this score, we gather statistics on the distribution of the classes on a set of images, and use them to quantify the severity of the bias. In the following, we investigate two distinct scenarios, namely *context-aware*, where we analyze the bias on caption-specific images $\mathcal{I}_b^t$, and *context-free*, where we consider the whole set of images $\mathcal{I}_b$ associated to one bias $b \in \mathcal{B}$.

3.2.1 Context-Aware Bias

As discussed in Section 1, our focus lies in examining bias exclusively when the classes are not explicitly mentioned in the caption. The bias proposals pipeline described in Sec. 3.1 filters out such cases; nevertheless, there could be additional aspects within the caption that impact the outcome. For example, the two captions *"A military is running"* and *"A person is running"* are both agnostic to the bias *"person gender"*, but the direction and magnitude of the bias may be very different in the two cases. To consider the role of the context in the bias assessment, we collect statistics at the caption level, analyzing the set of images $\mathcal{I}_b^t$ produced from a specific caption $t \in \mathcal{T}$. Given a bias b we compute the probability for a class $c \in \mathcal{C}_b$ as:

$$p(c|t, \mathcal{C}_b, \mathcal{D}_b) = \frac{1}{|\mathcal{I}_b^t|} \sum_{I \in \mathcal{I}_b^t} \mathbb{1}\big(\hat{c} = c\big) \tag{4}$$

with $\hat{c} = \texttt{VQA}(I, q, \mathcal{C}_b)$ the prediction of the VQA as defined in Eq. (3), and $\mathbb{1}(\cdot)$ the indicator function.

3.2.2 Context-Free Bias

Differently from the context-aware scenario, our interest lies in characterizing the overall behavior of the model G. This is crucial as it offers valuable insights into aspects such as the majority class (*i.e.* the direction toward which the bias tends) and the overall intensity of the bias. To effectively exclude the role of the context in the captions, we propose to average the VQA scores for $c \in \mathcal{C}_b$ over all captions t related to that bias $b \in \mathcal{B}$:

$$p(c|\mathcal{C}_b, \mathcal{D}_b) = \frac{1}{|\mathcal{D}_b|} \sum_{(t,q) \in \mathcal{D}_b} p(c|t, \mathcal{C}_b, \mathcal{D}_b) \tag{5}$$

Note that the context-aware bias is a special case of this scenario, where $\mathcal{D}_b$ has a single instance, *i.e.* $\mathcal{D}_b = \{(t, q)\}$.

3.2.3 Bias Quantification and Ranking

After collecting the scores for each individual attribute class $c \in \mathcal{C}_b$, we can aggregate them to rank the severity of biases within the generative model. As mentioned in Sec. 1, we follow existing work [15, 60] and consider the model G as unbiased with respect to a concept b when the distribution of the possible classes $c \in \mathcal{C}_b$ is uniform. To quantitatively assess the severity of the bias, we compute the entropy of the probability distribution of the classes obtained using either Eq. (4) or Eq. (5). To compare biases with different numbers of classes, we normalize the entropy by the maximum possible entropy [59]. Additionally, we adjust the score for enhanced human readability. In practice, our bias severity score is defined as follows:

$$\bar{\mathcal{H}}_b = 1 + \frac{\sum_{c \in \mathcal{C}_b} \log p(c|\mathcal{C}_b, \mathcal{D}_b)}{\log(|\mathcal{C}_b|)} \tag{6}$$

The resulting score is always bounded $\bar{\mathcal{H}}_b \in [0, 1]$, where 0 indicates an unbiased concept while 1 a biased one.

We note that, while we focused our pipeline on conditional generative models, our model can be easily extended for studying biases in both real-world multimodal datasets (*e.g.* by assuming images $\mathcal{I}_b^t$ are provided rather than generated), and to unconditional generative models (*i.e.* by using a captioning system on their outputs as set $\mathcal{T}$). We refer the reader to the *Supp. Mat.* for details where we will also show an analysis between the unconditional GAN *StyleGAN3* [28] and its training set FFHQ [27].

4. Experiments

In this section, we conduct a series of experiments to assess the proposed framework quantitatively. In Sec. 4.1, we provide implementation details and the preprocessing steps applied to the datasets. In Sec. 4.2, we quantitative evaluate OpenBias on two directions, (i) comparing it with a state-of-the-art classifier-based method on a closed set of well-known social biases, (ii) testing the agreement between OpenBias and human judgment via a user study.

4.1. Pipeline Implementation

Datasets. We study the bias in two multimodal datasets Flickr 30k [61] and COCO [33]. Flickr30k [61] comprises 30K images with 5 caption per image, depicting images in the wild. Similarly, COCO [33] is a large-scale dataset containing a diverse range of images that capture everyday scenes and objects in complex contexts. We filter this dataset, creating a subset of images whose caption contains a single person. This procedure results in roughly 123K captions. Our choice is motivated by building a large subset of captions specifically tied to people. This focus on the person-domain is crucial as it represents one of the most sensitive scenarios for exploring bias-related settings. Nevertheless, it is worth noting that the biases we discover within this context extend beyond person-related biases to include objects, animals, and actions associated with people. Further details are highlighted in Sec. 5.

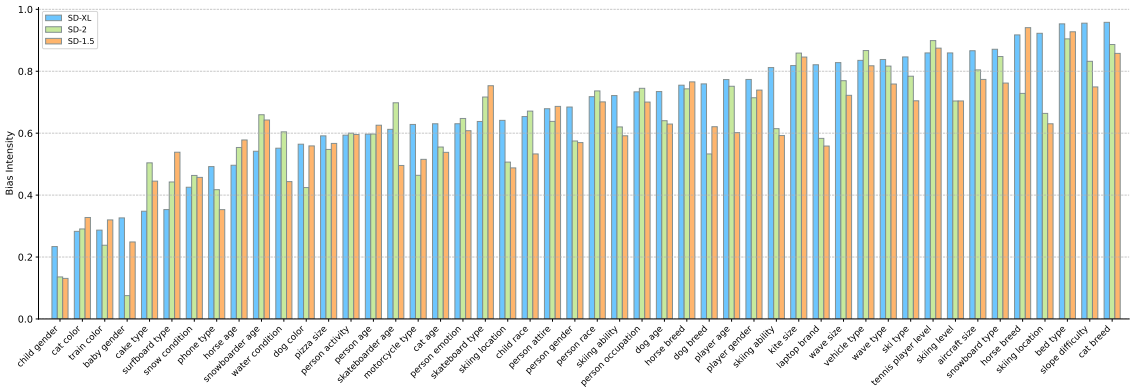

Figure 3. Comparison of context-aware discovered biases on Stable Diffusion XL, 2 and 1.5 [41, 44] with captions from COCO [33].

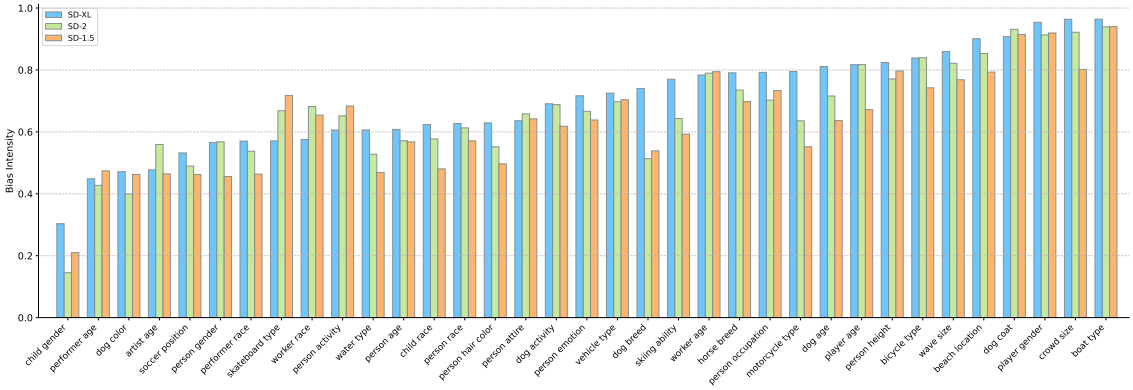

Figure 4. Comparison of context-aware found biases on Stable Diffusion XL, 2 and 1.5 [41, 44] on captions from Flick30k [61].

Implementation Details. Our pipeline is designed to be flexible and modular, enabling us to replace individual components as needed. In this study, we leverage LLama2-7B [54] as our foundation LLM. This model is exploited to build the knowledge base of possible biases, as described in Sec. 3.1. We refer the reader to the *Supp. Mat.* for details regarding the prompts and examples we use to instruct LLama to perform the desired tasks. To assess the presence of the bias, we rely on state-of-the-art Visual Question Answering (VQA) models. From our evaluation outlined in Sec. 4.2, Llava1.5-13B [34, 35] emerges as the top-performing, thus we adopt it as our default VQA model. Finally, we conduct our study by randomly selecting 100 captions associated with each bias and generating $N = 10$ images for each caption using a different random seed. In this way, we obtain a set of 1000 images, that we use to study the context-free and context-aware bias of the target generative model.

4.2. Quantitative Results

Our open-set setting harnesses the zero-shot performance of each component. As in [15], we evaluate OpenBias using FairFace [26], a well-established classifier fairly trained, as the ground truth on gender, age, and race. While FairFace treats socially sensitive attributes as closed-set, we uphold our commitment to inclusivity by also evaluating OpenBias with self-identified ones, reported in the *Supp. Mat.*.

Model	Gender		Age		Race	
	Acc.	F1	Acc.	F1	Acc.	F1
CLIP-L [42]	91.43	75.46	58.96	45.77	36.02	33.60
OFA-Large [56]	**93.03**	83.07	53.79	41.72	24.61	21.22
mPLUG-Large [31]	**93.03**	82.81	61.37	52.74	21.46	23.26
BLIP-Large [32]	92.23	82.18	48.61	31.29	36.22	35.52
Llava1.5-7B [34, 35]	92.03	82.33	66.54	62.16	55.71	42.80
Llava1.5-13B [34, 35]	92.83	**83.21**	**72.27**	**70.00**	**55.91**	**44.33**

Table 1. VQA evaluation on the generated images using COCO captions. We highlight in gray the chosen default VQA model.

Model	Flickr 30k [61]			COCO [33]		
	gender	age	race	gender	age	race
Real	0	0.032	0.030	0	0.041	0.028
SD-1.5 [44]	0.072	0.032	0.052	0.075	0.028	0.092
SD-2 [44]	0.036	0.069	0.047	0.060	0.045	0.105
SD-XL [41]	0.006	0.028	0.180	0.002	0.027	0.184

Table 2. KL divergence (↓) computed over the predictions of Llava1.5-13B and FairFace on generated and real images.

Agreement with FairFace. We compare the predictions of multiple SoTA Visual Question Answering models with FairFace. Firstly, we assess the zero-shot performance of the VQA models on synthetic images, performing our comparisons using images generated by SD XL. The evaluation involves assessing accuracy and F1 scores, which are computed against FairFace predictions treated as the ground truth. The results are reported in Tab. 1. Llava1.5-13B

emerges as the top-performing model across different tasks, consequently, we employ it as our default VQA model.

Next, we evaluate the agreement between Llava and FairFace [26] on different scenarios. Specifically, we run the two models on real and synthetic images generated with Stable Diffusion 1.5, 2, and XL. We measure the agreement between the two as the KL Divergence between the probability distributions obtained using the predictions of the respective model. We report the results in Tab. 2. We can observe that the models are highly aligned, obtaining low KL scores, proving the VQA model's robustness in both generative and real settings. *Supp. Mat.* provides a more comprehensive evaluation of the VQA.

User Study. We conduct a human evaluation of the proposed pipeline at the context-aware level, to assess its alignment with human judgment. The study presents 10 images generated from the same caption for each bias. We use public crowdsourcing platforms, without geographical restrictions, and randomizing the questions' order. Each participant is asked to identify the direction (majority class) of each bias and its intensity in a range from 0 to 10. The option *"No bias"* is provided to capture the instances where no bias is perceived, corresponding to a bias intensity of 0. We conduct the user study on a subsection of the biases, resulting in 15 diverse object-related and person-related biases and 390 diverse images. We collect answers from 55 unique users, for a total of 2200 valid responses. The user study results are shown in Fig. 5, where we compare the bias intensity as collected from the human participants with the severity score computed with OpenBias. We can observe that there is a high alignment on various biases such as *"Person age"*, *"Person gender"*, *"Vehicle type"*, *"Person emotion"* and *"Train color"*. We compute the Absolute Mean Error (AME) between the bias intensity produced by the model and the average user score, resulting in an AME $= 0.15$. Furthermore, we compute the agreement on the majority class, *i.e.* the direction of the bias. In this case, OpenBias matches the collected human choices 67% of the cases. We remark that concepts of bias and fairness are highly subjective, and this can introduce further errors in the evaluation process. Nevertheless, our results show a correlation between the scores, validating our pipeline.

5. Findings

In this section, we present our findings from the examination of three extensively utilized text-to-image generative models, specifically Stable Diffusion XL, 2, and 1.5 [41, 44]. We use captions from Flickr and COCO, as detailed in Sec. 4.1. We structure our findings by examining the biases of different models and delineating the distinctions between context-free and context-aware bias.

Rankings. We present here the biases identified by our pipeline on Stable Diffusion XL, 2, and 1.5 [41, 44], in

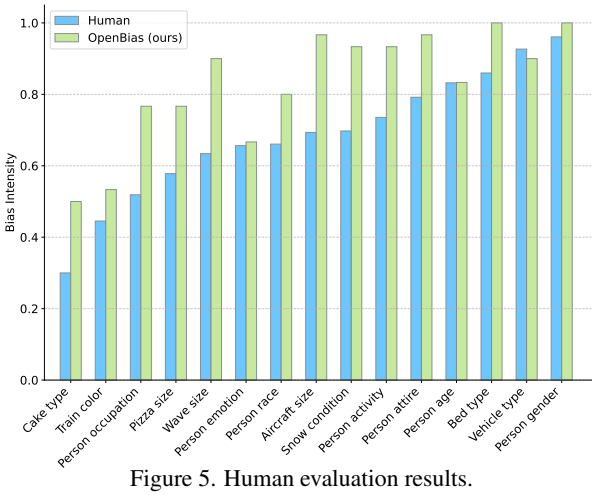

Figure 5. Human evaluation results.

Fig. 3 and 4. Importantly, OpenBias identifies both well-known (*e.g. "person gender"*, *"person race"*) and novel biases (*e.g. "cake type"*, *"bed type"* and *"laptop brand"*). From the comparison of different models, we observe a correlation between the intensities of the biases across different Stable Diffusion versions. We note, however, a subtle predominance of SD XL in the amplification of bias compared to earlier versions of the model. Moreover, the set of proposed biases varies depending on the initial set of captions used for the extraction. Generally, biases extracted from Flickr are more object-centric compared to those from COCO, aligning with the filtering operation applied to the latter. This difference highlights the potential of OpenBias to propose a tailored set of biases according to the captions it is applied to, making the bias proposals domain-specific.

Context-Free vs Context-Aware. Next, we study the different behavior of a given model, when compared in a context-free vs context-aware scenario (see Sec. 3 for formal definition). This analysis assesses the influence of other elements within the captions on the perpetuation of a particular bias. In Fig. 9 we report the results obtained on SD XL. It is noteworthy to observe that, in this case, the correlation between the scores is not consistently present. For example, the intensity score for *"motorcycle type"* is significantly higher when computed within the context, compared to the same evaluation free of context. This discrepancy suggests that there is no majority class (*i.e.* the general direction of the bias), but rather the model generates motorcycles of one specific type in a given context. Vice versa, for *"bed type"* we observe a high score in both settings, suggesting that the model always generates the same type of bed.

Qualitative Results. We show examples of biases discovered by OpenBias on Stable Diffusion XL. We present the results in a context-aware fashion and visualize images generated from the same caption where our pipeline identifies a bias. We organize the results in three sets and present unexplored biases on objects and animals, novel biases as-

Figure 6. Novel biases discovered on Stable Diffusion XL [41] by OpenBias.

Figure 7. Novel person-related biases identified on Stable Diffusion XL [41] by OpenBias.

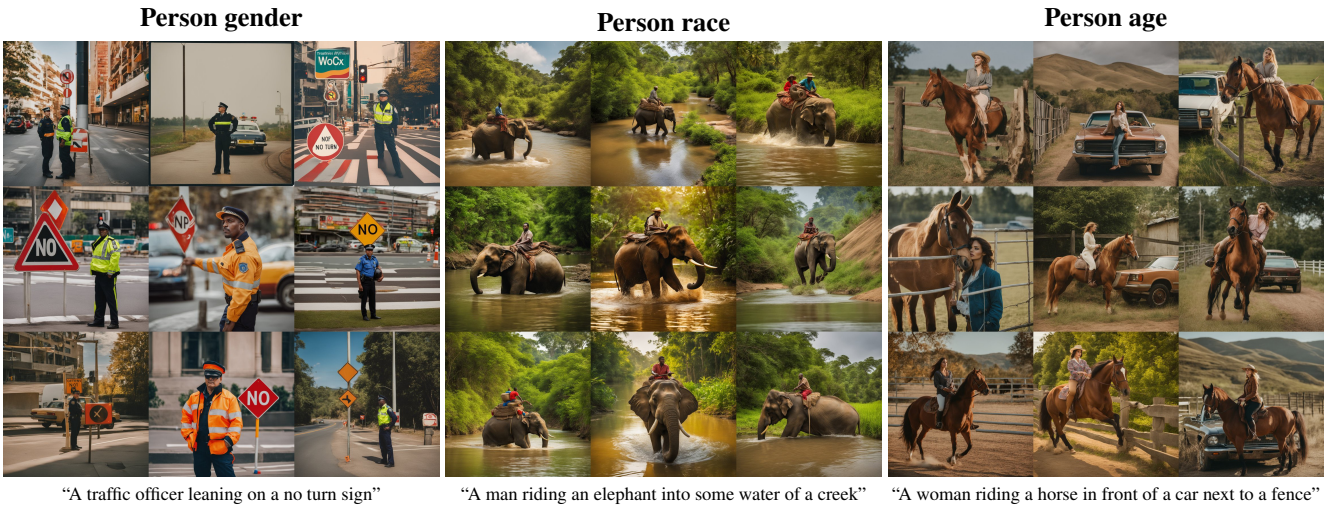

Figure 8. Person-related biases found on Stable Diffusion XL [41] by OpenBias.

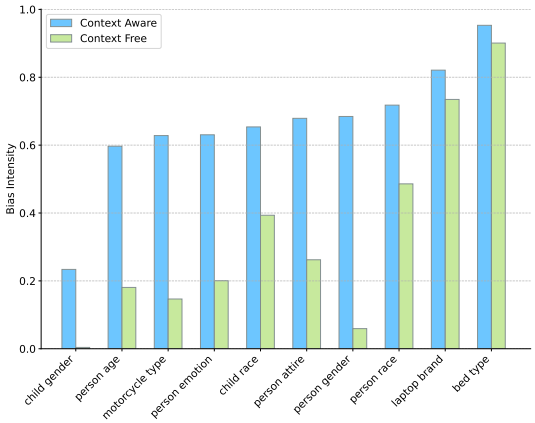

Figure 9. Highlighting the importance of the context aware approach on Stable Diffusion XL [41] on the captions from COCO.

sociated with persons, and well-known social biases. We highlight biases discovered on objects and animals in Fig. 6. For example, the model tends to generate *"yellow"* trains or *"quarter horses"* even if not specified in the caption. Furthermore, the model generates laptops featuring a distinct *"Apple"* logo, showing a bias toward the brand.

Next, we display novel biases related to persons discovered by OpenBias. For instance, we unveil unexplored biases such as the *"person attire"*, with the model often generating people in a formal outfit rather than more casual ones. Furthermore, we specifically study *"child gender"* and *"child race"* diverging from the typical examination centered on adults. For example, in Fig. 7 second column, we observe that the generative model links a black child with an economically disadvantaged environment described in the caption as *"a dirt road"*. The association between racial identity and socioeconomic status perpetuates harmful stereotypes and proves the need to consider novel biases within bias mitigation frameworks. Lastly, we show qualitative results on the well-studied and sensitive biases of *"person gender"*, *"race"*, and *"age"*. In the first column of Fig. 8, Stable Diffusion XL exclusively generates *"male"* officers, despite the presence of a gender-neutral job title. Moreover, it explicitly depicts a *"woman"* labeled as *"middle-aged"* when engaged in horseback riding. Finally, we observe a *"race"* bias, with depictions of solely black individuals for *"a man riding an elephant"*. This context-aware approach ensures a thorough comprehension of emerging biases in both novel and socially significant contexts. These results emphasize the necessity for more inclusive open-set bias detection frameworks. We provide additional qualitatives and comparisons in the *Supp. Mat.*.

6. Limitations

OpenBias is based on two foundation models to propose and quantify biases of a generative model, namely LLama [54] and LLava [35]. We rely on the prediction of these models, without considering their intrinsic limitations. Existing research [17, 38] highlights the presence of biases in these models which may be propagated in our pipeline. Nevertheless, the modular nature of our pipeline provides flexibility, allowing us to seamlessly incorporate improved models should they become available in the future. Finally, in this work, we delve into the distinction between context-free and context-aware biases, revealing different behaviors exhibited by models in these two scenarios. However, our evaluation of the role of the context is only qualitative. We identify the possibility of systematically studying the context's role as a promising future direction.

7. Conclusions

AI-generated content has seen rapid growth in the last few years, with the potential to become even more ubiquitous in society. While the usage of such models increases, characterizing the stereotypes perpetrated by the model becomes of significant importance. In this work, we propose to study the bias in generative models in a novel open-set scenario, paving the way to the discovery of biases previously unexplored. We propose OpenBias, an automatic bias detection pipeline, capable of discovering and quantifying traditional and novel biases without the need to pre-define them. The proposed method builds a domain-specific knowledge base of biases which are then assessed and quantified via Vision Question Answering. We validate OpenBias showing its agreement with classifier-based methods on a closed set of concepts and with human judgement through a user study. Our method can be plugged into existing bias mitigation works, extending their capabilities to novel biases. OpenBias can foster further research in open-set scenarios, moving beyond classical pre-defined biases and assessing generative models more comprehensively.

Ethical statement and broader impact. This work contributes to fairer and more inclusive AI, by detecting biases in T2I generative models. We conduct our research responsibly, transparently, and with a strong commitment to ethical principles. Despite this, due to technical constraints, socially sensitive attributes, such as gender, are treated as closed sets for research purposes only. Moreover, OpenBias entails the biases of the LLM and VQA models, thus it may not discover all possible biases. *We do not intend to discriminate against any social group but raise awareness on the challenges of detecting biases beyond closed sets.*

Acknowledgments: This work was supported by the MUR PNRR project FAIR (PE00000013) funded by the NextGenerationEU and by the EU Horizon projects ELIAS (No. 101120237) and AI4Media (No. 951911), NSF CAREER Award #2239840, and the National AI Institute for Exceptional Education (Award #2229873) by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education, and Picsart AI Research (PAIR).

References

[1] Sharat Agarwal, Sumanyu Muku, Saket Anand, and Chetan Arora. Does data repair lead to fair models? curating contextually fair data to reduce model bias. In *WACV*, 2022. 1

[2] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *CVPR*, 2023. 1

[3] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*, 2016. 1

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint*, 2021. 2

[5] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. In *NeurIPS*, 2023. 2

[6] Manuel Brack, Felix Friedrich, Patrick Schramowski, and Kristian Kersting. Mitigating inappropriateness in image generation: Can there be value in reflecting the world's ugliness? *arXiv preprint*, 2023. 2

[7] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 1

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 2, 3

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2

[10] Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. Video chatcaptioner: Towards the enriched spatiotemporal descriptions. *arXiv preprint*, 2023. 2

[11] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*, 2023. 1

[12] Moreno D'Incà, Christos Tzelepis, Ioannis Patras, and Nicu Sebe. Improving fairness using vision-language driven image augmentation. In *WACV*, 2024. 1

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[14] Dave Epstein, Allan Jabri, Ben Poole, Alexei A Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *NeurIPS*, 2023. 1

[15] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint*, 2023. 1, 2, 4, 5

[16] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint*, 2022. 1

[17] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *arXiv preprint*, 2023. 8

[18] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Xingqian Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: A comprehensive multimodal object-level image editor. *arXiv preprint*, 2023. 1

[19] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *CVPR*, 2023. 2

[20] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 1

[21] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2022. 1

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2

[23] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023. 2

[24] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. In *ICML*, 2023. 1

[25] Sangwon Jung, Sanghyuk Chun, and Taesup Moon. Learning fair classifiers with partially annotated group labels. In *CVPR*, 2022. 1

[26] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *WACV*, 2021. 5, 6

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 4

[28] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 4

[29] Patrik Joslin Kenfack, Kamil Sabbagh, Adín Ramírez Rivera, and Adil Khan. Repfair-gan: Mitigating representation bias in gans using gradient clipping. *arXiv preprint*, 2022. 1, 2

[30] Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Bias-to-text: Debias-

ing unknown visual biases through language interpretation. *arXiv preprint*, 2023. 2

[31] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. 5

[32] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 5

[33] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 4, 5

[34] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. 5

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 5, 8

[36] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023. 1

[37] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *NeurIPS*, 2020. 1

[38] Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: Origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 2023. 8

[39] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, 2022. 1

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. In *Transactions on Machine Learning Research*, 2023. 2

[41] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 1, 2, 5, 6, 7, 8

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 5

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint*, 2022. 1

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 5, 6

[45] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 1

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 1

[47] Yash Savani, Colin White, and Naveen Sundar Govindarajulu. Intra-processing methods for debiasing neural networks. In *NeurIPS*, 2020. 1

[48] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, 2017. 3

[49] Hongjin SU, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Selective annotation makes language models better few-shot learners. In *ICLR*, 2023. 3

[50] Xingzhe Su, Yi Ren, Wenwen Qiang, Zeen Song, Hang Gao, Fengge Wu, and Changwen Zheng. Unbiased image synthesis via manifold-driven sampling in diffusion models. *arXiv preprint*, 2023. 2

[51] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. 2

[52] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *ICCV*, 2023. 2

[53] Shuhan Tan, Yujun Shen, and Bolei Zhou. Improving the fairness of deep generative models without retraining. *arXiv preprint*, 2021. 2

[54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint*, 2023. 2, 5, 8

[55] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, 2018. 1

[56] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022. 5

[57] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *CVPR*, 2020. 1

[58] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. 2

[59] Allen R Wilcox. Indices of qualitative variation. Technical report, Oak Ridge National Lab., Tenn., 1967. 4

[60] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018. 1, 4

[61] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2014. 4, 5

[62] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Itigen: Inclusive text-to-image generation. In *ICCV*, 2023. 1, 2

[63] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1

[64] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017. 1

[65] Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. In *Transactions on Machine Learning Research*, 2023. 2